SYNOPTIC NOTE

Recognition and Classification of Medical Texts using Convolutional Neural Network

Title of research : Recognition and Classification of Medical Texts using Convolutional Neural Network
Course and subject : Ph.D in Computer Science
Name of student : Miss. Nayan Jadhav
Enrolment date : 16/12/2019
Department : Karnataka State Akkamahadevi Women's University
Department of Computer Science, vijaypur
Name of the guide : Dr. Aziz Makandar (Professor & Chairperson)
Karnataka State Akkamahadevi Women's University
Department of Computer Science, vijaypur

Introduction

There is a need for effective and efficient literature mining and knowledge discovery that can help biologists to gather and make use of the knowledge encoded in text documents. In order to make organized and structured information available, automatically recognizing biomedical entity names becomes critical and is important for information retrieval, information extraction and automated knowledge acquisition.

Descriptive naming convention: Many biomedical entity names are descriptive, e.g. 'normal thymic epithelial cells'. Such phenomenon highlights the difficulty for identifying the left boundaries of biomedical entity names.

Conjunction and disjunction: Two or more biomedical entity names may share one head noun by using conjunction or disjunction construction. For ex ample, '91 and 84 kDa proteins'

Non-standardized naming convention: One biomedical entity name may be found with various spelling forms, e.g. 'Nacetylcysteine', 'N-acetyl-cysteine', 'NAcetylCysteine', etc
Abbreviation: Abbreviations are frequently used in the biomedical domain. Many abbreviations in the biomedical domain are formed quite irregularly. For example, abbreviations can be from a subset of syllable boundaries, e.g. 'IL2

Medical text recognition system identifies different entities in many ways such name of person locations, and organizations from news articles, reports, blogs, tweets. Main steps of named entity recognition are boundary detection of entities and classification of entities into already defined classes. This results of recognition and classification is widely used in information retrieval and extraction.

The main component of biomedical natural language processing is named entity recognition system which extracts information from the text and finally does the knowledge discovery. As amount of health and biomedical text being available is huge and since much of the data is recorded in non-structured text, like in clinical notes and biomedical publications the bottleneck of biomedical information processing is how to make use of the knowledge resources and build scalable models to process large amounts of text. Medical text recognition, also known as biomedical concept identification or concept mapping, is a key step in biomedical language processing.

One major focus of TM research has been on medical text Recognition , a crucial initial step in information extraction, aimed at identifying chunks of text that refer to specific entities of interest, such as gene, protein, drug and disease names. Medical text recognition aims to find entities in biomedical texts, an invaluable function that becomes very important for further processing such as information retrieval, information extraction and knowledge discovery.

The medical text is natural language processing is used for finding medical related name entity from the given datasets it has 2 sub tasks are as follows

a) Recognition: This is used to identify the medical related names in the given datasets.
b) Normalization: this is known as grounding which is to say identified name with its meaning.

Literature review

A medical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set

Ahmed Sultan Al-Hegami et al., proposed that This paper gives the comparison between different subsets of features and three classification approach (Naïve Bayes, K-Nearest Neighbour and decision tree) for biomedical named entity recognition. The proposed model uses Support Vector Machine(SVM) and Naive Bayes algorithms which are satisfactory and effective for BNER. More time required to train complex algorithm.Does not perform well on multiple class tasks.

A Comparative Study of Segment Representation for medical Named Entity Recognition
H.L.Shashirekha et al., proposed the model with the three phases- Tokenization, Boundary detection, Type classification. Support Vector Machines (SVMs) and Conditional Random fields (CRFs) are used to train different BioNER models used in this paper. The performance of f-score degrades with the complexity of model. Use limited size of context rather on whole text. Affected by data distribution.

Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts

Shaodian Zhang et al., gives possible improvements on the approach including nested NPs as candidates, better chunker for medical text, better domain representations, and improved IDF values of phrases. The paper gives solution to tackle the challenges of entity boundary detection and entity type classification.

Recognizing names in biomedical texts: a machine learning approach

GuoDong Zhou et al., In order to make organized and structured information available; automatically recognizing biomedical entity names becomes critical and is important for information retrieval, information extraction and automated knowledge acquisition.

Drug Name Recognition in medical Texts: Feature Conjunction and Feature Selection

Shengyu Liu et al., Drug name recognition (DNR) is a critical step for drug information extraction. Machine learning-based methods have been widely used for DNR with various types of features such as part-of-speech, word shape, and dictionary feature. Features used in current machine learning-based methods are usually singleton features which may be due to explosive features and a large number of noisy features when singleton features are combined into conjunction features.

Medical Named Entity Recognition Using Machine Learning Algorithms

Dr. K.S. Wagh et al., Named-entity recognition system (NER) [1] identifies different entities in many ways such name of person locations, and organizations from news articles, reports, blogs, tweets. Main steps of named entity recognition are boundary detection of entities and classification of entities into already defined classes. This results of recognition and classification is widely used in information retrieval and extraction.

BioBERT: a pre-trained biomedical language representation model for biomedical text mining

jinhyuk Lee et al., Biomedical text mining is becoming increasingly important as the number of biomedical documents rapidly grows. With the progress in natural language processing (NLP), extracting valuable information from biomedical literature has gained popularity among researchers, and deep learning has boosted the development of effective biomedical text mining models

Named Entity Recognition and Relation Detection for medical information extraction

nadeesha Perera et al., In this paper, we reviewed methods for Named Entity Recognition (NER) and Relation Detection (RD) allowing, e.g., to identify interactions between proteins and drugs

or genes and diseases. Over the years, many methods have been introduced and studied for resolving a variety of problems in biomedical, health, and clinical sciences.

Drug name recognition in biomedical Drug name recognition in medical method

Linna He et al., there is an urgent need to develop a technology for extracting drug information automatically from biomedical texts, and drug name recognition is an essential prerequisite for extracting drug information. This article presents a machine-learning-based approach to recognize drug names in biomedical texts.

Named entity recognition on bio-medical literature documents using hybrid based approach

R.Ramachandran et al., The proposed hybrid approach identified named entity and outperformed well than the existing baseline method. Many organizations are involving in retrieving the hidden information from the literature documents. Extracting the drug names, diseases, symptoms, route of administration, species and dosage forms from the textual document is an easy task due to the innovation of technologies in the Natural Language Processing.

Medical named entity recognition using deep neural networks with contextual information

Hyejin Cho et al., In biomedical text mining, named entity recognition (NER) is an important task used to extract information from biomedical articles. Previously proposed methods for NER are dictionary- or rule-based methods and machine learning approaches. However, these traditional approaches are heavily reliant on large-scale dictionaries, target-specific rules, or well-constructed corpora

Improving biomedical named entity recognition with syntactic information

Yuanhe Tian et al., Biomedical named entity recognition (BioNER) is an important task for understanding biomedical texts, which can be challenging due to the lack of large-scale labeled training data and domain knowledge. To address the challenge, in addition to using powerful encoders (e.g., biLSTM and BioBERT), one possible method is to leverage extra knowledge that is easy to obtain.

Named Entity Recognition using Machine Learning and pattern Selection Rules

Sanjana Kamath et al., discusses about significance of machine learning in NER. The paper has proposed methods and rules namely hybrid method and maximum entropy model. The data used are extracted from tagged data sets. This model focuses on unidentified words. The paper mainly focuses on entropy model and this work can be shifted to any other domain as the data is trained data set.

Character-level neural network for medical named entity recognition

Mourad gridach et al., Biomedical named entity recognition (BNER), which extracts important named entities such as genes and proteins, is a challenging task in automated systems that mine knowledge in biomedical texts. The previous state-of-the-art systems required large amounts of task-specific knowledge in the form of feature engineering, lexicons and data pre-processing to achieve high performance.

Problem statement

Medical Texts Recognition is a natural language processing task that attempts to find named entities (persons, locations, organizations, etc.) within a text. Since the tag of one word depends on neighboring words (and their tags), the problem is modeled as a sequence labeling task, wherein a sentence is considered as an input, and the goal is to produce the tags of all words jointly, so as to maximize accuracy.

Motivation

Medical Texts is a crucial preprocessing task in Biomedical Natural Language Processing pipeline as it is a prerequisite for extracting huge amount of high-valued biomedical knowledge deposited in unstructured text .Traditional techniques which are applied for newswire documents cannot be used directly to extract Medical due to the complexity. Usually single classifiers are used to classify the Medical Texts. But, a single classifier may perform well on some data and may not give good results for some other data.

Challenges of Medical text recognition

- Ambiguity: same word or phrase referring to different entities according to the context (e.g. myc-c refers to gene name as well as protein name)
- Variations: multiple spelling forms for same Medical Texts (e.g. RNase protein or RNaseP and intergin alpha 4 or alpha4 intergin).
- Synonymy: different names refer to same entity (e.g. caspase-3, CASP-3, apoptosis-related cysteine protease, and CPP32 denote the same entity).
- Abbreviations: same word or abbreviation interpreted as different definitions (e.g. ACE can be either Angiotensin Converting Enzyme or Affinity Capillary Electrophoresis).
- Multi-word: A Medical Texts having multiple words (e.g. CD28 surface receptor).
- Nested Medical Texts: one Medical text occurring within longer Medical text (e.g. T cell is nested within nuclear factor of activated T cells family protein).
- Growing Vocabulary: a major problem for effective Medical Texts models is the overwhelming growth rate of newly added Medical Texts, which always results in Out-of-Vocabulary.

Drawbacks of existing

- The problem is that current evaluation practices and resources in Medical texts do not allow us to decide.
- Ambiguity and Abbreviations -One of the major challenges in identifying named entities is language. Recognizing words which can have multiple meanings or words that can be a part of different sentences
- Spelling Variations-The English language plays a very important role. Words which do not make a major difference in phonetics but make a major difference in the way of writing and its spelling
- Foreign Words-Words which are not used very frequently these days, or words that are not heard by a lot of people, is another major challenge in this field. Words like person names, location names etc.

Applications

- Gene cluster identification
- Protein interactions
- Gene-disease associations
- Gene-trait associations
- Protein-disease associations
- NER is used effectively for automatic identification of events like disasters and crimes.
- Medical Entity Recognition refers to a sub task of information extraction that locates entities in medical field (E.g. Medicine names, Medical tools etc.)
- Major in this field is the complicated terms (medicine names) used and hence makes it difficult for identifying.

Research features

- Identification of the medication names given to patients.
- Identification of related information about dosages, duration, frequency, mode of administration, reason for prescription.
- Includes extraction of biomedical entities using Machine Learning and text mining which include Name, Problem and Test from Lab Reports and mapping appropriate Treatment using widely accepted datasets.
- Improve the performance by investigating more on conjunction and disjunction construction, the synonym phenomenon, the cascaded entity name phenomenon and inclusion of a reasonable biomedical dictionary.
- It will be enabling automatic generation of informative features to extract new and unknown characteristics of entity names.

- Enriching the dictionaries by adding more objects will give more accuracy. The work will be extended to update the dictionary from the suggestions of domain expert and retrain the model instantly.
- We would like to develop unsupervised learning methods for the Medical texts
- Medical text tools are only available for text reports, the Medical Texts can also be implemented for the previously implemented Medical texts tools are only available for text reports, the Medical Text can also be implemented for images, by extracting information by scanning the reports such as X-Rays, ECG

Research objective

- By using existing datasets on Medical text database creation on various features(genes, diseases, treatments, proteins,)
- Medical text recognition system identifies different entities in many ways such name of person locations, and organizations from news articles, reports, blogs, and tweets.
- Transformation of input data into a set of features. Using features like sentence splitting, tokenization, etc
- The predictions will be performed based on the information that they encode.
- Recognizing biomedical related texts and providing that text with particular meaning.
- Providing the necessary information regarding the recognized texts.

Methodology

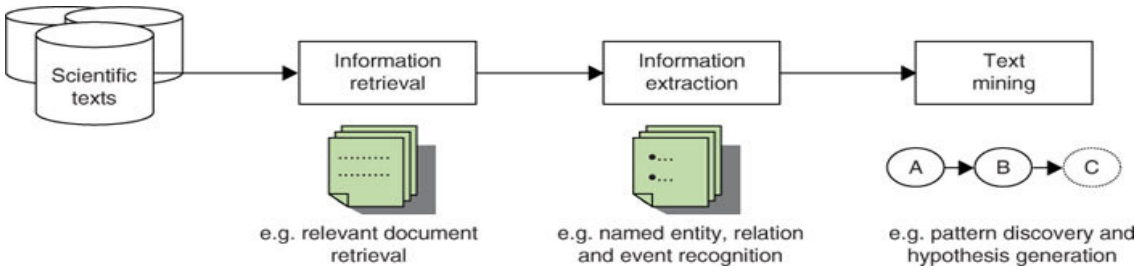

Fig: Block diagram of biomedical name recognition

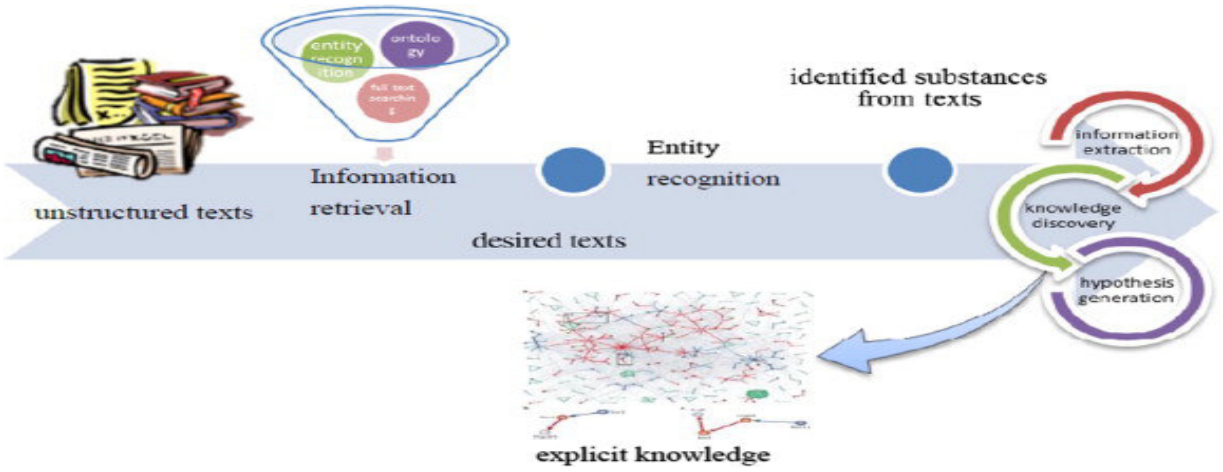

Fig: Block diagram of biomedical name recognition using text mining

In our proposing work we are using techniques such as conditional random fields (CRF), Medical Text recognition, Preprocessing, feature extraction, supervised learning, unsupervised learning, text mining, Document classification and clustering, Relationship discovery, Information extraction, Information retrieval, boundary detection, entity recognition, BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining), Convolutional Neural Network(CNN)

Proposing method

The system is evaluated using 2 datasets i.e. 2b2 and GENIA corpora. The i2b2 corpus contains clinical notes which contain Problems, Tests, and Treatments which are annotated entities. GENIA corpus is a collection of biomedical related words which contains biological entities such as DNA, RNA, and protein.

A general overview of Medical Text system. It generally consists of phases namely

1. Preprocessing
2. Feature extraction
3. Feature processing
4. Entity Classification
5. Text mining

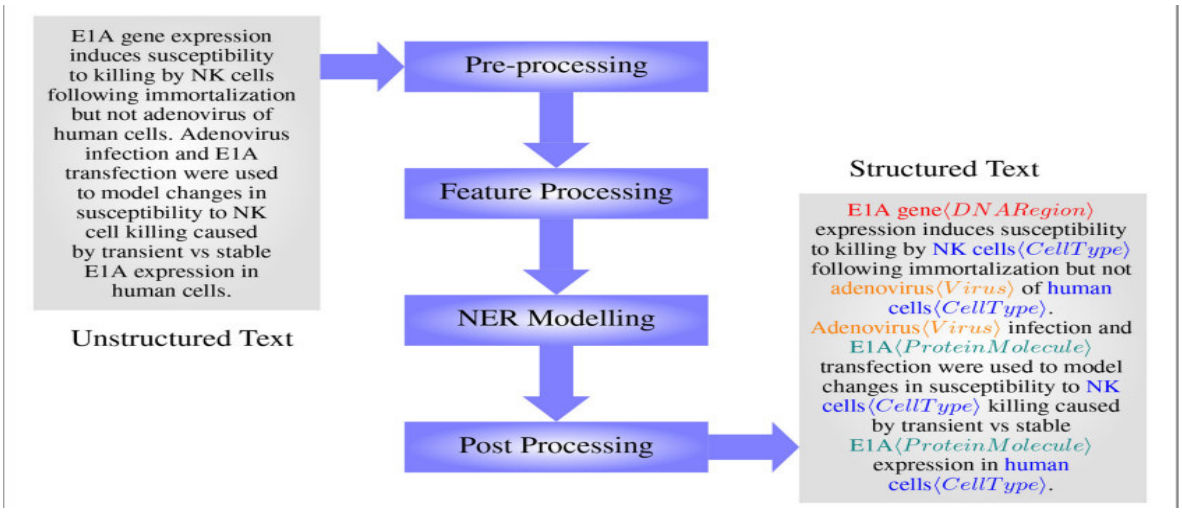
E1A gene expression induces susceptibility to killing by NK cells following immortalization but not adenovirus of human cells. Adenovirus infection and E1A transfection were used to model changes in susceptibility to NK cell killing caused by transient vs stable E1A expression in human cells.
Unstructured Text
Pre-processing
Feature Processing
NER Modelling
Post Processing
Structured Text
E1A gene⟨DNARegion⟩ expression induces susceptibility to killing by NK cells⟨CellType⟩ following immortalization but not adenovirus⟨Virus⟩ of human cells⟨CellType⟩. Adenovirus⟨Virus⟩ infection and E1A⟨ProteinMolecule⟩ transfection were used to model changes in susceptibility to NK cells⟨CellType⟩ killing caused by transient vs stable E1A⟨ProteinMolecule⟩ expression in human cells⟨CellType⟩.

Fig: proposing method for Bio-NER

1. Pre-processing: Natural Language Processing (NLP) and text mining solutions can be accomplished by computerized systems in an effective manner. However, it is necessary firstly to properly delimit the documents into meaningful units.
2. Feature Extraction: Transformation of input data into a set of features. Features are distinctive properties of input patterns that help in differentiating between the categories of input patterns
    a) Sentence splitting: Sentence splitting is the process of breaking a text document into its respective sentences. In the end, each sentence should provide a specific local, logical and meaningful context for future tasks.
    b) Tokenization: Tokenization is the process of breaking a sentence into its constituent meaningful units, called tokens.
    c) Annotation encoding: In order to internally represent the annotated entity names, it is necessary to use an encoding scheme to give a tag to each token of the text.
3. Feature processing: Feature processing is a crucial NER task, since the predictions will be performed based on the information that they encode, reflecting special phenomena and linguistic characteristics of the naming conventions.
    a) Orthographic: The purpose of orthographic features is to capture knowledge about word formation. For example, a word that starts with a capital letter could indicate the occurrence of an entity name (e.g. in the protein name “MyoD”).
    b) Morphological: Morphological features, on the other hand, reflect common structures and/or sub-sequences of characters among several entity names, thus identifying similarities between distinct tokens.

c) Context: Higher-level relations between tokens and extracted features can be established through windows or conjunctions of features, reflecting the local context of each token.

4. Entity classification: Entity classification is done by using signature generation method. The similarity based method is mainly used in word sense disambiguation (WSD), which assumes that meaning of a word is closely related to the distribution of different words surrounding it.
5. Text mining: Extraction of useful and non-trivial knowledge from texts. Collecting and linking information based on databases or knowledge with information extracted from texts.

Conclusion

Medical Text Recognition and Relation Detection (RD) allowing e.g., to identify interactions between proteins and drugs or genes and diseases .Over the years, many methods have been introduced and studied for resolving a variety of problems in biomedical, health, and clinical sciences. For this reason, we aimed for a systematic presentation by categorizing methods according to their main characteristics. Biomedical entity recognition is an important task that is still unresolved but can help in other medical-related systems. For example, the NER can identify important findings that are essential for safe and effective health care. The NER task applied to the biomedical domain aims to extract and identify entities of interest that can be used by health professionals.

References

[1] Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat (2008), Named Entity Recognition Approaches.

[2] Weiyi Sun, Anna Rumshisky, Ozlem Uzuner (2013), Evaluating temporal relations in clinical text: 2012 i2b2 Challenge.

[3] J.-D. Kim, T. Ohta, Y.Tateisi, J. Tsujii (2003), GENIA corpus—a semantically annotated corpus for bio-textmining.

[4] Anish Talwar, Yogesh Kumar(2013), Machine Learning: An artificial intelligence methodology.

[5] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, Parve Kuksa (2011), Natural Language Processing (ALmost) from Scratch.

[6] Ahmed Sultan Al-Hegami,Ameen Mohammed Farea Othman,Fuad Tarbosh Bagash,(2017). A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set.

[7] Almas Tasneem,Archana B,(2016). A Survey on Biomedical Named Entity Extraction.

[8] S.Vijaya, Dr.R.Radha,(2017).Biomedical Named Entity Recognition - a swift review.

[9] Samet Atdağ and Vincent Labatut. A Comparison of Named Entity Recognition Tools Applied to Biographical Texts.

[10] H. L. Shashirekha,Hamada A. Nayel,(2016) A Comparative Study of Segment Representation for Biomedical Named Entity Recognition.
[11] Mohammed RAIS,Abdelmonaime LACHKAR,Said EL ALAOUI OUATIK(2014),A Comparative Study of Biomedical Named Entity Recognition Methods Based Machine Learning Approach.
[12] Shaodian Zhang , Noémie Elhadad(2013),Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts.
[13] Dr. K.S. Wagh1 , Aishwarya Kulkarni2 , Pratiksha Pawar3 , Neha Kirange4 , Shraddha Kashid(2017), Bio Medical Named Entity Recognition Using Machine Learning Algorithms
[14] David Campos, Sérgio Matos and José Luís Oliveira(2016), Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools
[15] R. Ramachandran1 , K. Arutchelvan1(2021), Named entity recognition on bio-medical literature documents using hybrid based approach